Team 4:

Computational and infrastructure environment in the next 5 years and the limits on high resolution initialized simulations

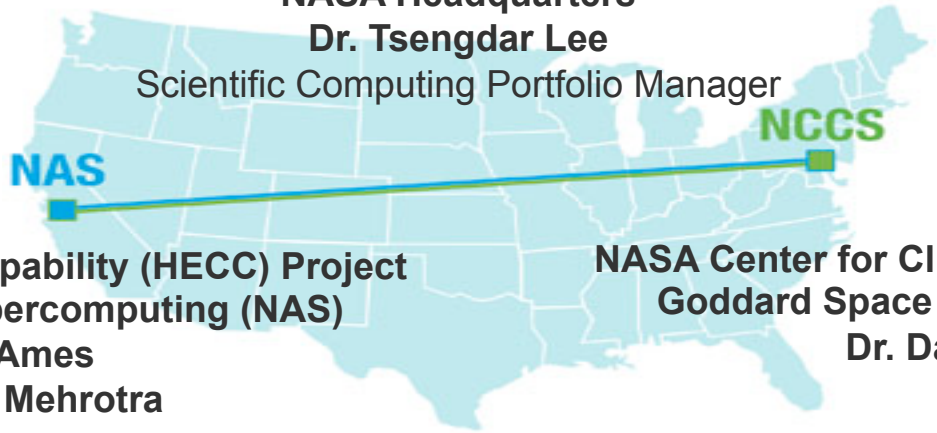
Team Leads: Bill Putman (NASA), Brian Gross (NOAA) and Bill Collins (DOE)

Additional input from:

Daniel Duffy - High Performance Computing Lead NASA Center for Climate Simulations

NASA High-End Computing Program

HEC Program Office
NASA Headquarters
Dr. Tsengdar Lee
Scientific Computing Portfolio Manager



High-End Computing Capability (HECC) Project
NASA Advanced Supercomputing (NAS)
NASA Ames
Dr. Piyush Mehrotra

NASA Center for Climate Simulation (NCCS)
Goddard Space Flight Center (GSFC)
Dr. Daniel Duffy

Current Major NCCS Services

Advanced Data Analytics Platform (ADAPT)

- Web services
- Designed for large scale data analytics
- Science Cloud
- HPC Technologies

- 1,000's of cores
- Petabytes of storage
- Using decommissioned HPC systems

Mass Storage

Tape ~45 PB

Disk ~4 PB

NCCS Local Area Network
10 GbE and 40 GbE

High Performance Computing Cluster Discover

- Large Scale Models
- ~3.5 PF Peak

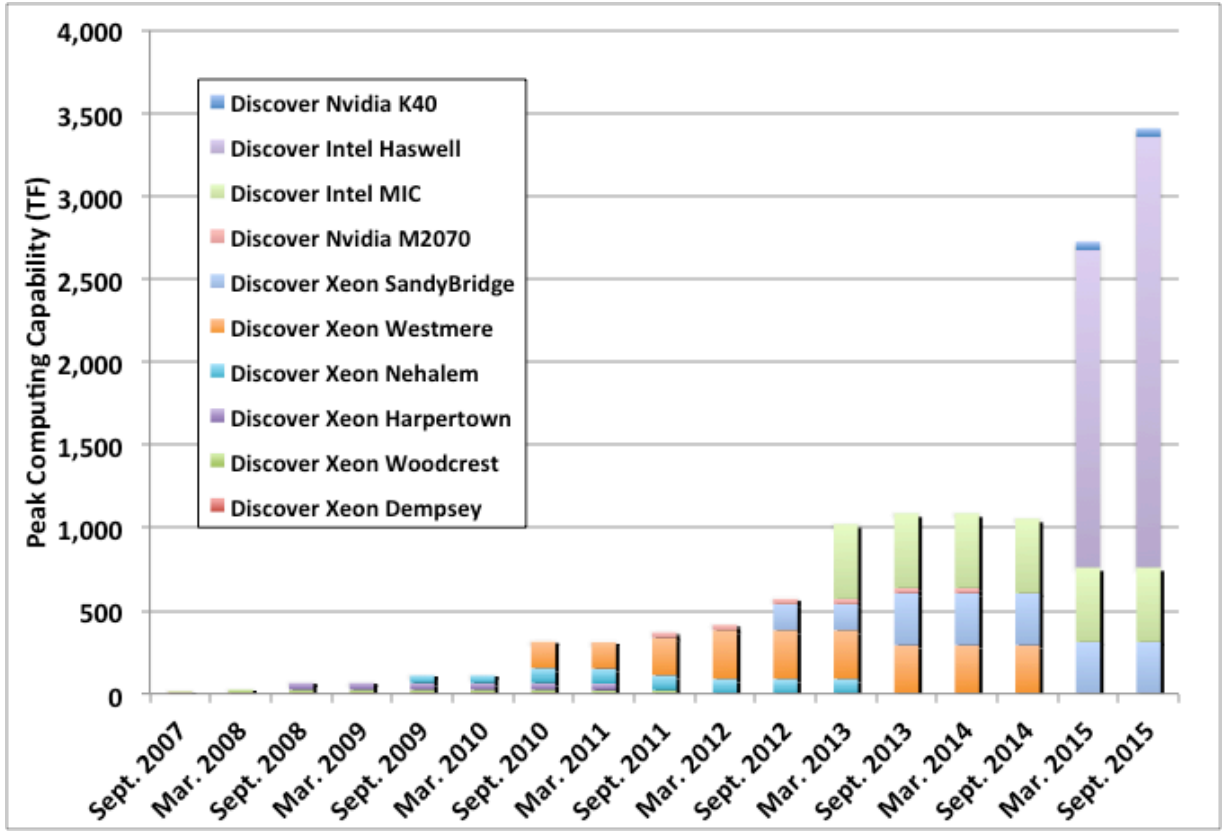
>80,000 Cores

GPFS Shared File System ~33PB

Current NCCS Total Peak Computing Capability

This graph shows the growth of the total peak computing capability of Intel Xeon processors, GPUs, and Intel Phis at the NCCS

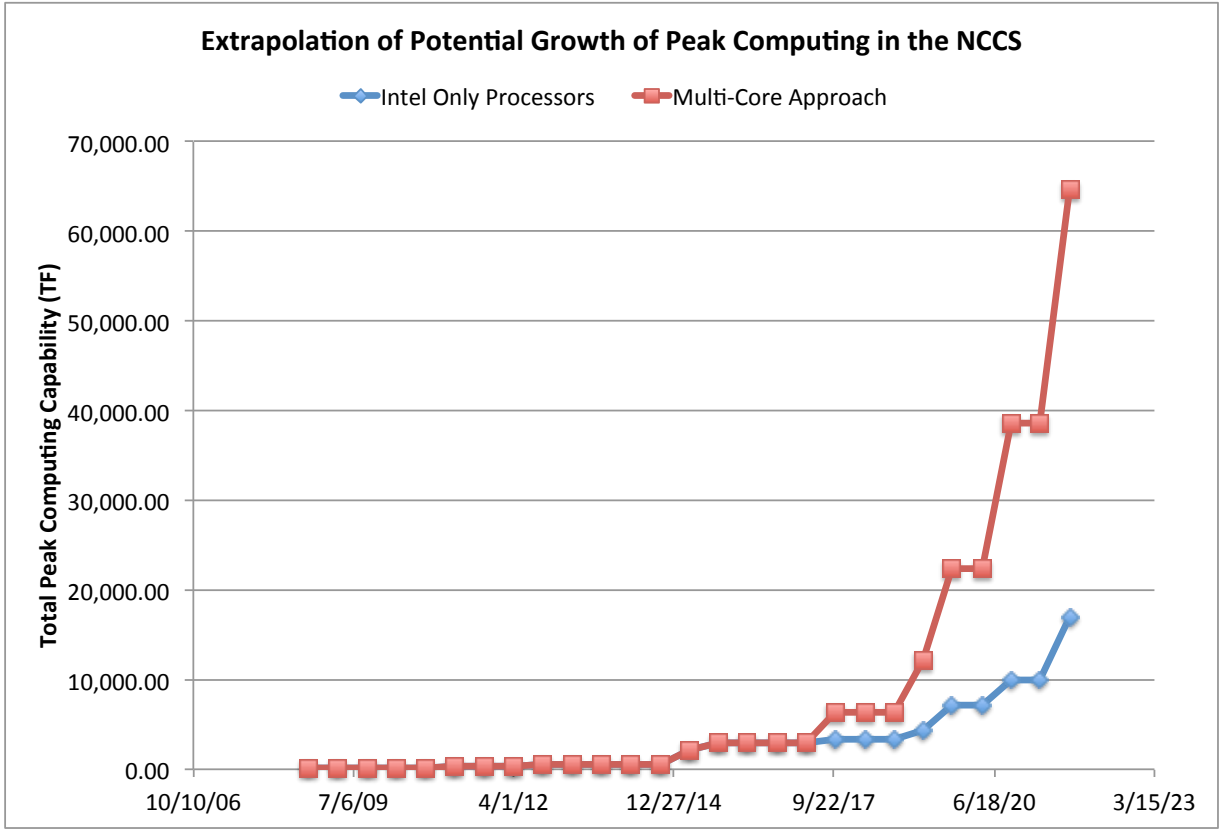
- Grew by almost 3x over the past few years
- Over a 100x increase in compute capability over the past 8 years
- Comparable growth in spinning disk over the same time period



Potential Growth of Peak Computing

Estimated expansion of capability at the NCCS assuming a continued Moore's Law growth over the next 5 years

- Blue line assumes only traditional Intel Xeon Processors ***O(100K) cores***
- Red line shows a multi-core approach something similar to NVIDIA GPUs or Intel Phis ***O(10M) cores***
- Over a ***3x increase in the peak*** computing capability using the multi-core approach and potentially a ***100x increase in the number of cores***
- Don't focus on the absolute numbers but on the relative increases.



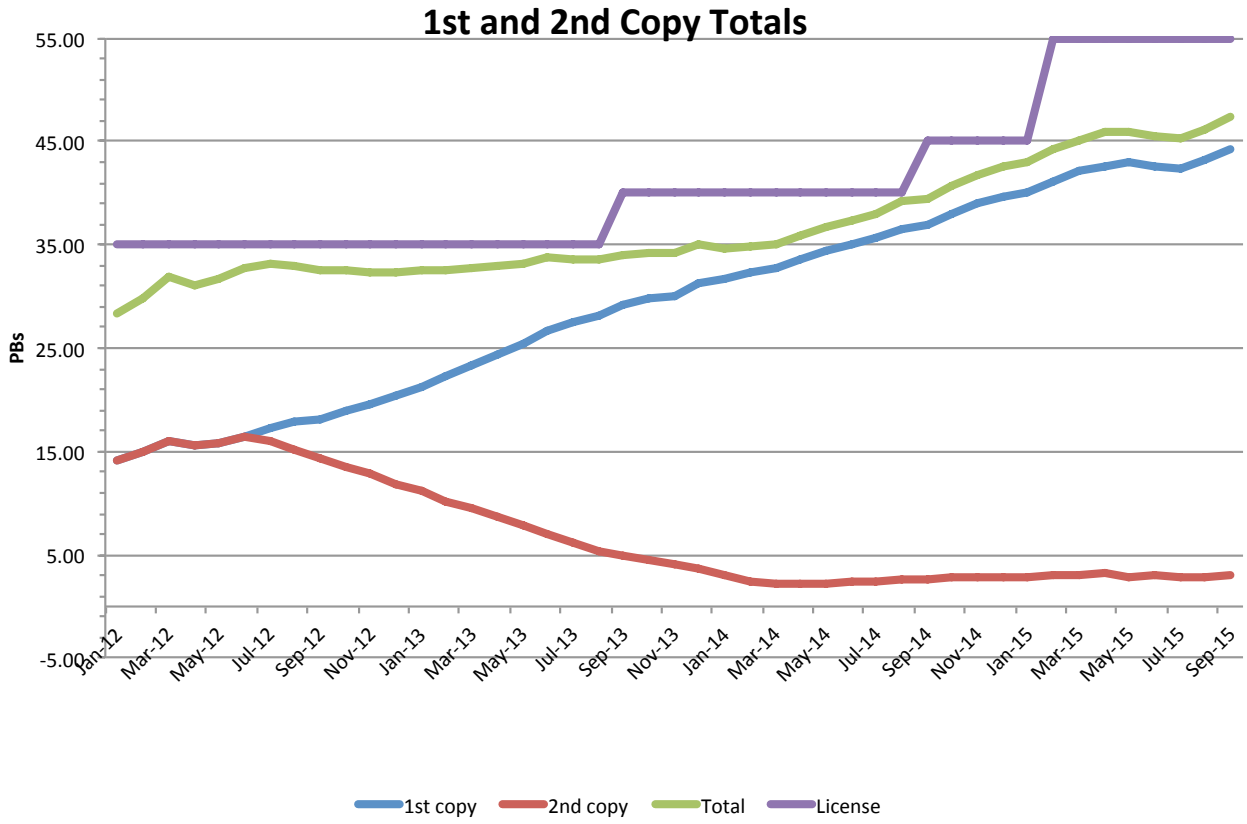
NCCS Discover Scratch (Disk)

Calendar	Description	Decommission	Total Usable Capacity (TB)
2012	Combination of DDN disks	None	3,960
Fall 2012	NetApp1: 1,800 by 3 TB Disk Drives; 5,400 TB RAW	None	9,360
Fall 2013	NetApp2: 1,800 by 4 TB Disk Drives; 7,200 TB RAW	None	16,560
Early 2015	DDN10: 1,680 by 6 TB Disk Drives, 10,080 TB RAW	DDNs 3, 4, 5	26,000
Mid 2015	DDN11: 1,680 by 6 TB Disk Drives, 10,080 TB RAW	DDNs 7, 8, 9	33,000
Mid 2016	Target additional 10,000 TB RAW	None	~41,000

- Usable capacity differs from RAW capacity for two reasons. First, the NCCS uses RAID6 (double parity) to protect against drive failures. This incurs a 20% overhead for the disk capacity. Second, the file system formatting is estimated to also need about 5% of the overall disk capacity. The total reduction from the RAW capacity to usable space is about 25%.

Growth of the NCCS Mass Storage (Tape)

- Stopped making second copies of tape in 2012.
- Close to linear growth regardless of compute resources.
- Quotas have been established as of 2015.
- No sign of this slowing down.



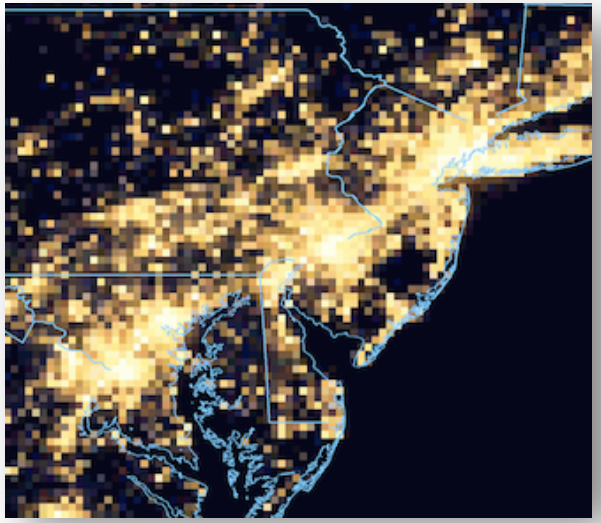
Typical GEOS-5 Output with Serial Writes

*Assumes IO Bandwidth of ~1 GB/s for 1-day simulation

7-km GEOS-5 7,200 cores	
0.06	GB per 2D Slice
4.45	GB per 3D Field *(72-levels)*
8.30	TB per Simulated Day
2.3	Hours Spent Writing*

3.5-km GEOS-5 14,400 cores	
0.25	GB per 2D Slice
17.8	GB per 3D Field *(72-levels)*
33.2	TB per Simulated Day
9.2	Hours Spent Writing*

1.25-km GEOS-5 43,200 cores	
1.0	GB per 2D Slice
71.2	GB per 3D Field *(72-levels)*
132.8	TB per Simulated Day
36.8	Hours Spent Writing*



Parallel IO Strategies

File-per-processor

Each processor maintains its own filehandle to a unique file and can write independently without any coordination with other processors.

Parallel file systems often perform well with this type of access up to several thousand files, but synchronizing metadata for a large collection of files introduces a bottleneck.

One way to mitigate this is to use a 'square-root' file layout, for example by dividing 10,000 files into 100 subdirectories of 100 files.

Shared-file (independent)

Shared file access allows many processors to share a common filehandle but write independently to exlusive regions of the file.

This coordination can take place at the parallel file system level.

However, there can be significant overhead for write patterns where the regions in the file may be contested by two or more processors.

Shared-file with collective buffering

Collective buffering is a technique used to improve the performance of shared-file access by offloading some of the coordination work from the file system to the application.

A subset of the processors is chosen to be the 'aggregators' who collect the data from all other processors and pack it into contiguous buffers in memory that are then written to the file system.

Reducing the number of processors that interact with the I/O subservers is often beneficial, because it reduces contention.

Asynchronous reads/writes

A collection of processors are reserved as IO servers, gathering data and distributing to designated writers or reading data and scattering back to the compute nodes.

In-Line data analysis

Perform some or all of your data anaylsis during processing to avoid having to write full resolution data to disk

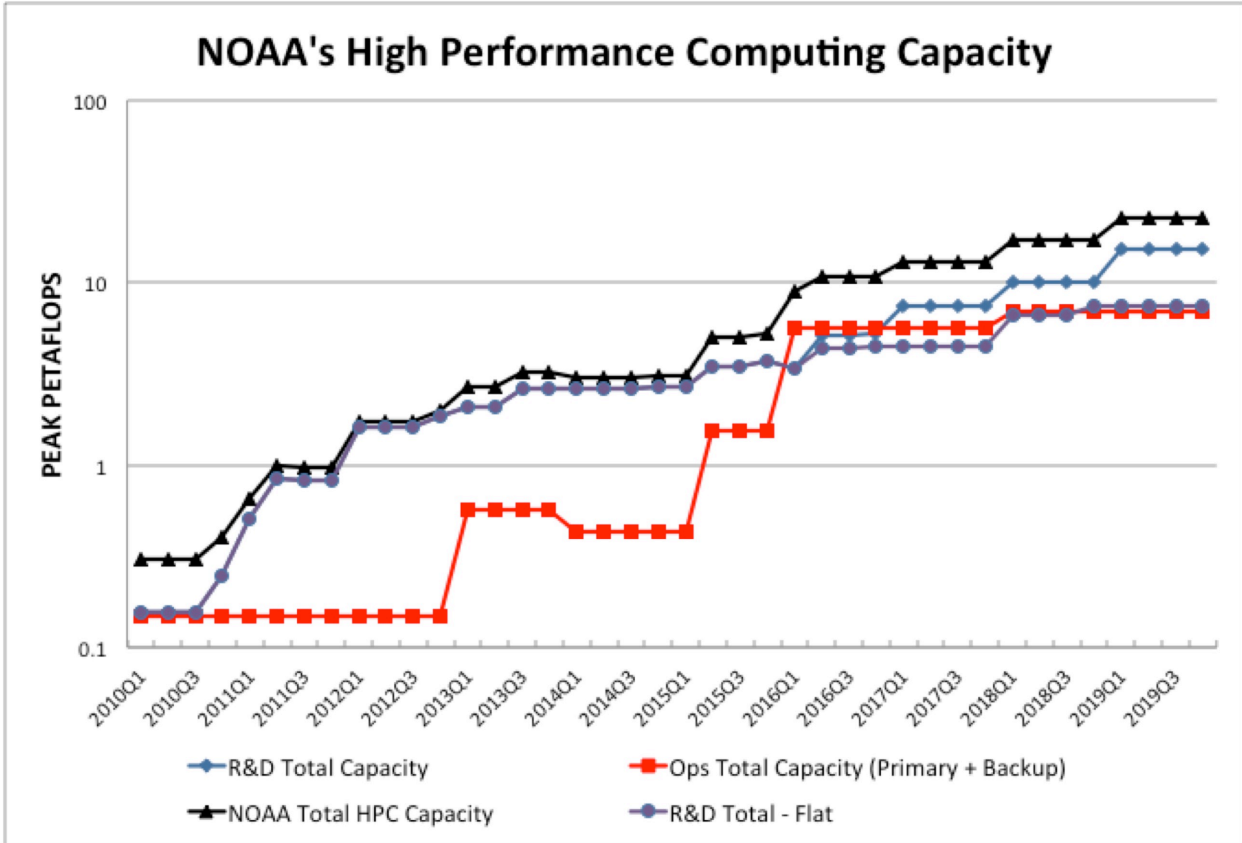
Background

NOAA High Performance Computing

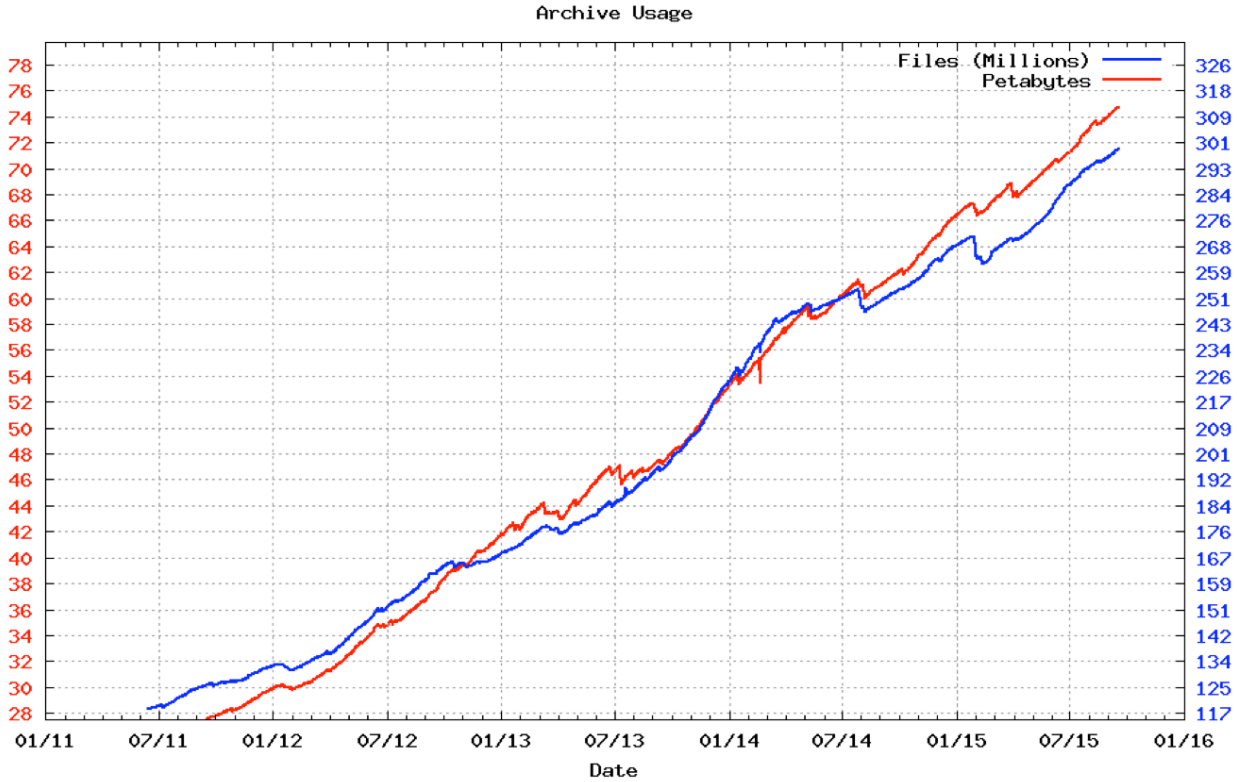
High Performance Computing System Investment

- Enables advancements in NOAA's environmental models, including assimilation of observational data and the generation of products to support mission services
- Budget supports supercomputers and associated facilities, networking, hardware/software maintenance, and systems support
- Resource made available and utilized by all NOAA line offices
- Operated by NOAA-wide integrated team

NOAA HPC Overview



Data Archive Growth



NOAA HPC Overview

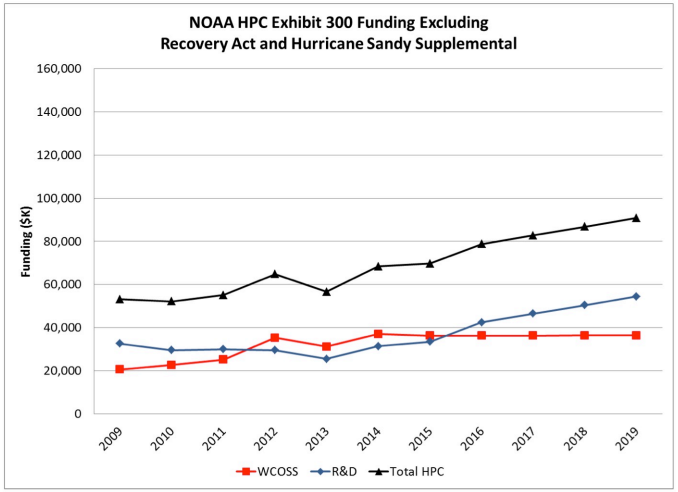


Chart includes FY16 PB increase profile for R&D HPCS

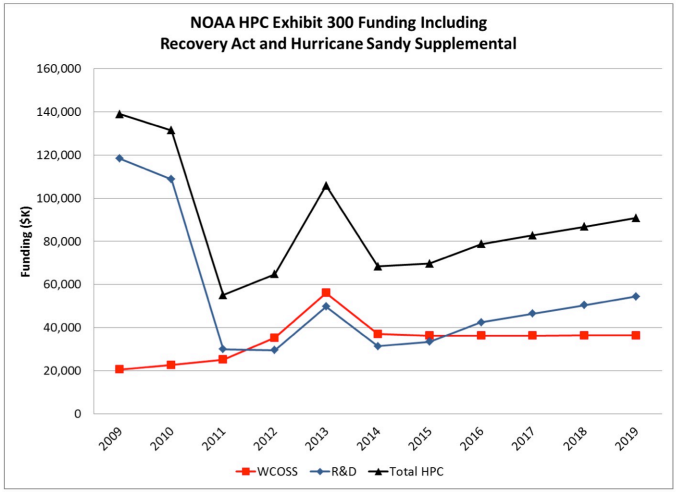


Chart includes FY16 PB increases profile for R&D HPCS

On software development ….

- Experience to date with fine-grained architectures: kernels can sing (~40X), but complex multi-physics codes croak (~<2X)
- Approach: code revisions for performance on conventional architectures will get us a significant way toward performance on fine-grained systems.
  - Component Concurrency
  - Offload I/O, Diagnostics
  - Performance analysis tools
  - vectorization (requires interaction with compiler vendors)
  - wide halos (to reduce comms)
  - nonmalleable executables (aka static memory)
  - direct use of coarray

National Strategic Computing Initiative

- Executive Order signed July 29, 2015 for a multi-agency strategic vision and Federal investment strategy in high-performance computing
- Objectives:
  - Accelerate delivery of a capable exascale computing system
  - Connect computing used for modeling and simulation to data analytic computing
  - Establish, over the next 15 years, a viable path forward for future HPC systems post-Moore's Law
  - Increase the capacity and capability of an enduring national HPC ecosystem
  - Develop public-private collaboration to share benefits between the US Gov't and industrial and academic sectors.
- Roles and Responsibilities
  - Lead Agencies: DOE, DOD, NSF
  - Foundational R&D Agencies: IARPA, NIST
  - Deployment Agencies: NASA, NOAA, NIH, FBI, DHS
    - These will develop mission-based HPC requirements to influence the early stages of the design of new HPC systems and will seek viewpoints from the private sector and academia on target HPC requirements
- https://www.whitehouse.gov/blog/2015/07/29/advancing-us-leadership-high-performance-computing

Questions to address in this workshop

Computing Capabilities

- What are the current computing capabilities available to NASA and NOAA and planned growth?

- The National Strategic Computing Initiative names DOE and NSF as Lead Agencies in getting us to exascale, and NASA and NOAA as deployment agencies that focus on their respective missions. Describe how these agencies can work together to advance initialized predictions.

- How do evolving multicore accelerator architectures fit into you current software models and to what extent have these technologies been explored to facilitate your hi-resolution coupled systems?

Resolution, Complexity, Ensembles

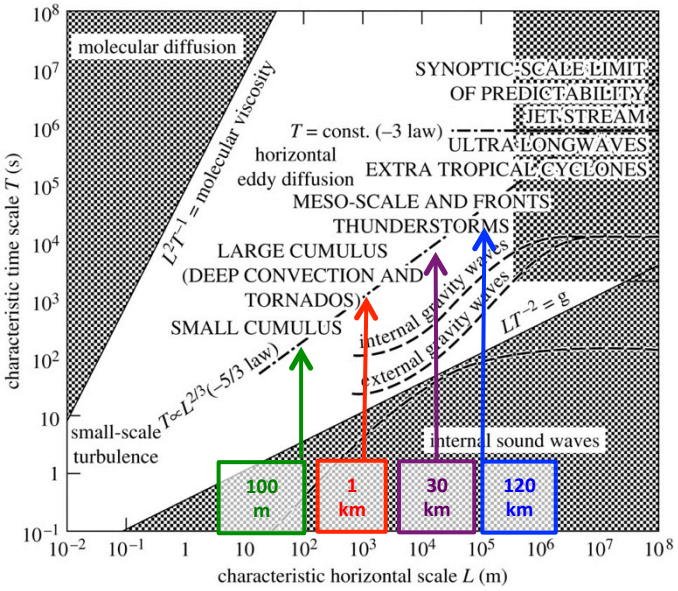
- What modeling improvements will most significantly impact computing and storage requirements (e.g., resolution, processes/complexity, ensemble members, etc)?
- What horizontal and vertical resolutions are necessary to adequately resolve processes in the coupled system that drive both prediction error in the short term forecast and climate simulation bias?
- What is the ideal size of the ensemble needed for this effort both for prediction and for understanding coupled processes and biases?

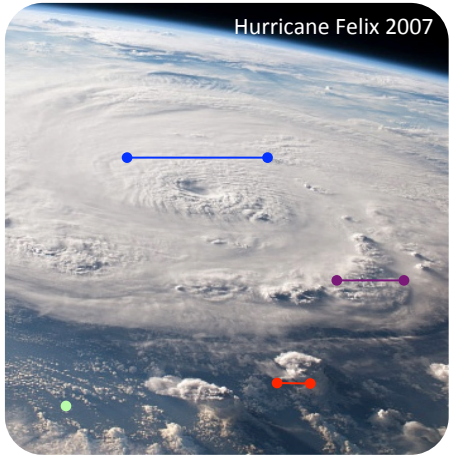
Data Volume/Storage/Sharing

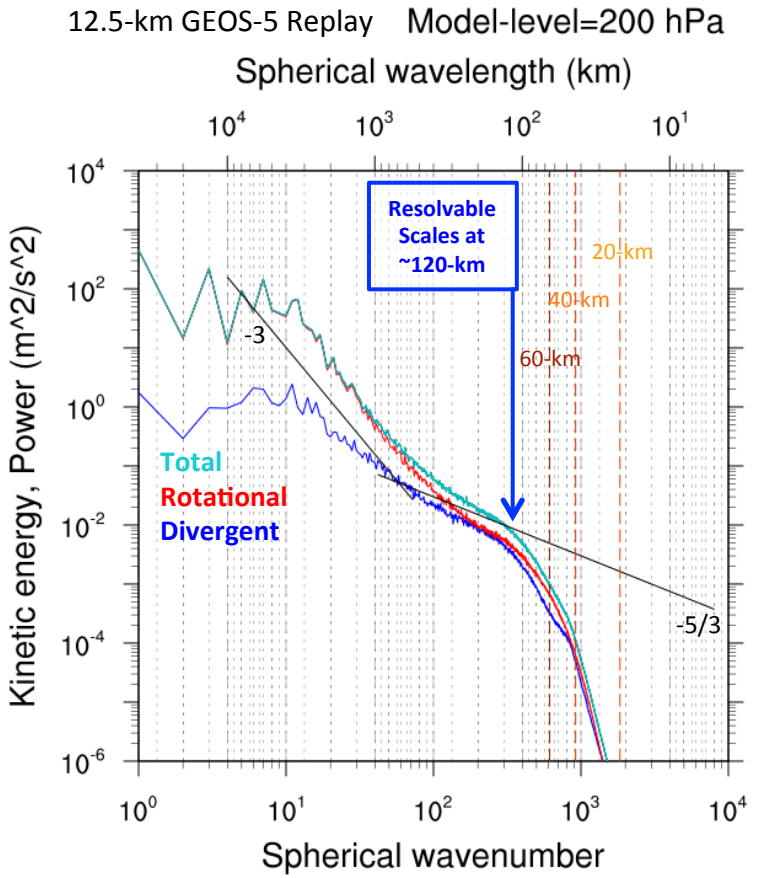
- How will increasingly high-resolution data be stored and shared for community research?
- As resolution increases, it becomes more difficult to save every bit to disk. How would you suggest we reduce the storage burden from coupled hi-res integrations?
- What must be analyzed at full resolution, and what can be evaluated at coarser spatial resolution?
- What aspects of your analysis can be in-lined during computation to reduce the required storage?
- How does increased horizontal resolution impact the necessary temporal resolution of your analysis and data storage?
- What new technologies, such as non-volatile random-access memory (NVRAM), provide the greatest potential to improve the scalability and efficiency of your coupled systems and particularly IO bottlenecks that are inevitable at high resolution?

Thank you!

Science and Computing Required to Increase Resolvable Scales in the Atmosphere

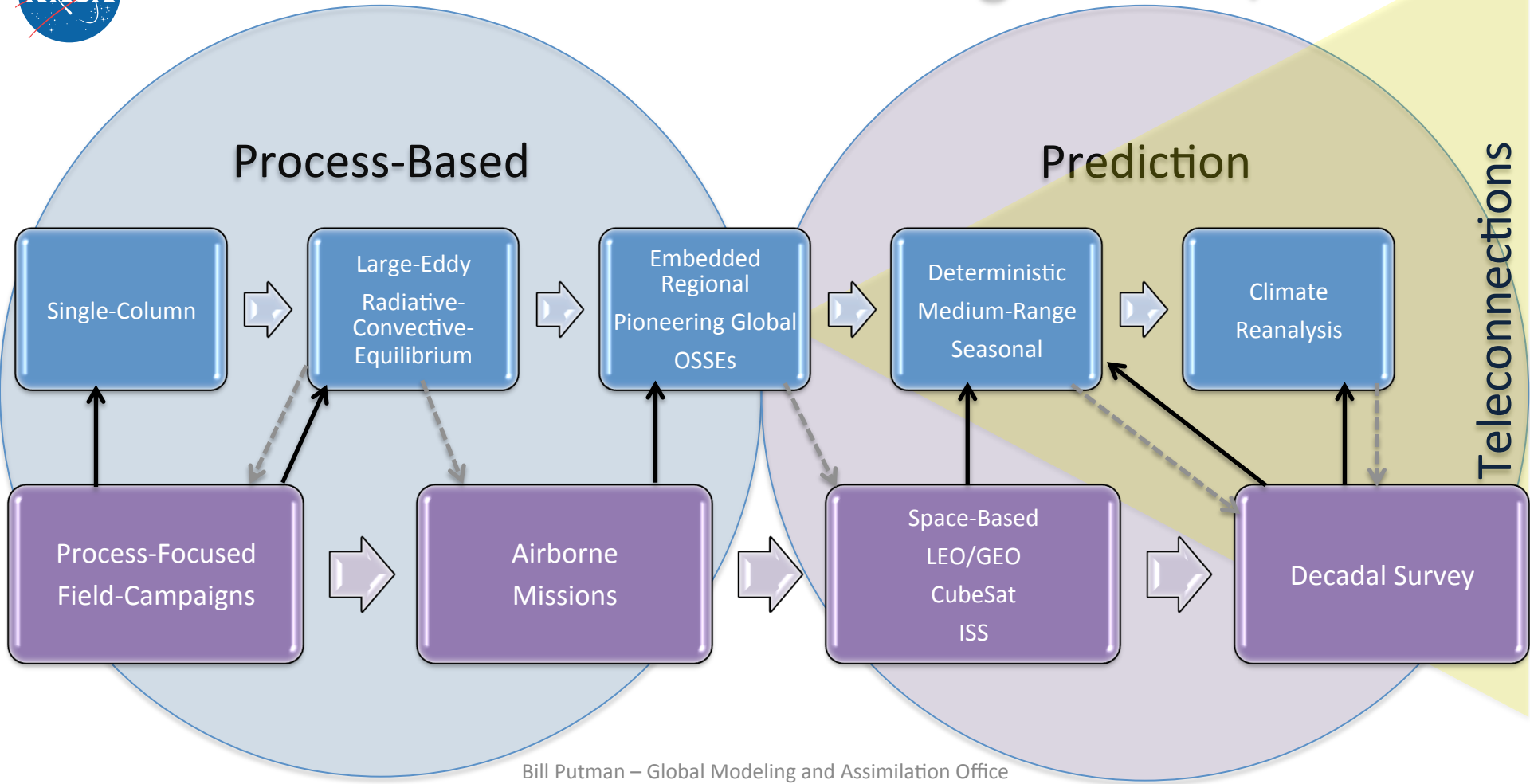






Resolution (km)	Resolvable ~10x (km)	Computing (Intel Based Cores)
25.0	250	800
12.5	**125**	**6,400**
3.0	**30**	**462,963**
1.0	10	12,500,000
0.1	**1**	**6,400,000,000**
10 (m)	**100 (m)**	**21,600,000,000,000,000**

Observation Driven Modeling at NASA/GMAO



Observation Driven Modeling at NASA/GMAO

Modeling Component	Current Capability	5-10 Years	10-20 Years
Single Column	100 Levels	200 Levels	400 Levels
Large-Eddy Radiative-Convective-Equilibrium	100 m - 1 km	10-100 m	1-10 m
Embedded-Regional Pioneering Global OSSEs	1-3 km	500 m – 1 km	100-500 m
Deterministic Medium-Range	10-15 km	1-3 km	500 m - 1 km
Seasonal	25 km	3-10 km	1-3 km
Climate/Reanalysis	50 km	10-25 km	3-10 km

Observation Driven Modeling at NASA/GMAO

Modeling Component	Current Processing (Intel-based cores)	5-10 Years (Intel-based cores)	10-20 Years (Intel-based cores)
Single Column	1	1	1
Large-Eddy Radiative-Convective-Equilibrium	1,000	500,000	1,000,000
Embedded-Regional Pioneering Global OSSEs	30,000	12,500,000	6,400,000,000
Deterministic Medium-Range	6,000	6,000,000	12,500,000
Seasonal	800	75,000	3,000,000
Climate/Reanalysis	100	8,000	100,000